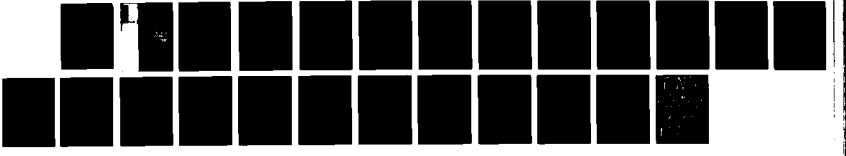
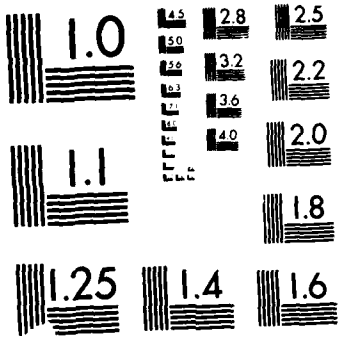
AD-A207 537 EXTENDED PINCUS THEOREMS AND CONVERGENCE OF SIMULATED ANNEALING(U) TEXAS UNIV AT AUSTIN CENTER FOR CYBERNETIC STUDIES A CHARNES ET AL. DEC 88 CCS-RR-615 1/1

UNCLASSIFIED DAKF15-87-C-0110 F/G 12/3 NL





UTION TEST CHART

(4)

CCS Research Report 615

EXTENDED PINCUS THEOREMS AND CONVERGENCE OF SIMULATED ANNEALING

by

A. Charnes
M. Wolfe*

CENTER FOR CYBERNETIC STUDIES
The University of Texas
Austin, Texas 78712

4

CCS Research Report 615

EXTENDED PINCUS THEOREMS AND CONVERGENCE OF SIMULATED ANNEALING

by

A. Charnes
M. Wolfe*

December 1988

*West Virginia University

The research for this paper was partly supported by National Science Foundation Grants SES 8520806 and SES 8722504 and U.S. Army Contract DAKF 15-87-C-0110 with the Center for Cybernetic Studies at The University of Texas at Austin. Reproduction in whole or in part is permitted for any purpose of the U.S. Government.

CENTER FOR CYBERNETIC STUDIES

A. Charnes, Director

College of Business Administration, 5.202
The University of Texas at Austin
Austin, Texas 78712-1177
(512) 471-1821

EXTENDED PINCUS THEOREMS AND CONVERGENCE OF SIMULATED ANNEALING

Abstract

Pincus' 1968 formula for the (unique) global minimum of a continuous function on a compact set in E^n is extended to finite multiple optima and to discrete and special variants. The impact of these on associated ergodic irreducible *aperiodic* Markov chain computation he initiated, (1970) currently called "simulated annealing", is exemplified and assessed leading to grave concern about what current simulated annealing processes may converge to instead of optima.

Keywords

Extended Pincus Theorems
Global Optimization
Simulated Annealing
Ergodic Markov Chains

EXTENDED PINCUS THEOREMS AND CONVERGENCE OF SIMULATED ANNEALING

Introduction

In 1968, M. Pincus [1] derived a closed form expression for the (assumed) unique global minimum point of a continuous function on a compact set S in E^n as the limit of the expected value (vector) of a one parameter family of distributions based on the continuous function F(x). Since these expressions involve integrations over irregular sets and in the many variables (of E^n), he established in 1970 [2] an approximate computational method based on the Metropolis-Ulam, von Neumann, et. al method [3] of associating an irreducible aperiodic Markov chain with the approximations to Riemann integrals and using the strong law of large numbers for such chains to generate an approximation to the unique global solution point. The relevance of this work to so-called "simulated annealing" as introduced in 1983 by Kirkpatrick, Gelatt and Vecchi [4] was not noticed by these authors since there is no mention of the Pincus work therein. Nor is there any reference to it in current 1988 work of Hajek [5] or by Chiang and Chow [6].

The analytic sharpness of the Pincus type formula is therefore here extended to situations of finite global optima, in several variants, thereby providing insight into what simulated annealing or sample Markov chain computations may converge to, both for finite discrete optimization problems and for continuum problems of non-convex type which Pincus seemed to have in mind.

The basic results are that the limit vector in Pincus-type formulas is a convex combination of the global minimization vectors. The convex proportions depend on the relative function shapes and volumes in the immediate neighborhoods of the individual global minima.

In the following we present an extended Pincus theorem and an extended discrete Pincus theorem as two extremes of a possible general Pincus formula which would involve a sum

of multiple integrals of different dimensions corresponding to different parts of the domain S of the function F whose global optima are sought. We proceed to examine for these implications of facts like the values of the function F are only known as one performs sample jumps on Markov chains whose transition probabilities are determined only as one makes sample jumps from a pre-selected irreducible symmetric aperiodic Markov transition matrix.

Such facts raise considerable concern about what it is that a simulated annealing process may converge to instead of a global optimum of the original problem. Indeed, our analytic expressions give additional support to the lack of assurance simulated annealers have in their results as exemplified by "re-annealing" and "annealing schedule" change devices introduced in the hope of obtaining "better" optima.

Extended Pincus Theorems

In [1] M. Pincus gave a closed form expression for the unique global minimum vector of a continuous function $F(x)$ over a compact set S in E^n as the limit of a sequence of expected value vectors. We here extend this result to the case of finite global minima for two extreme cases, one in which S is a discrete set of points. The proof arguments involve breaking the integrals or sums involved into two parts, one in the neighborhood of a global optimum and the other negligible for large parameter values.

Theorem 1:

Let (1) $F(x)$ be continuous over a compact set $S \subseteq E^n$

(2) The global minima z^α, $\alpha = 1, \ldots, k$ of $F(x)$ are contained in disjoint closures $\bar{0}_\alpha$ of open sets 0_α, with $\bar{\theta}_\alpha \subseteq S$.

Then there exists $\{\lambda_r\}$, $0 \le \lambda_r \to \infty$ and $0 \le \mu_\alpha$, $\sum_{\alpha=1}^{k} \mu_\alpha = 1$

such that

$$\lim_{\lambda_r \to \infty} \int_S x_i \exp\left(-\lambda_r F(x)\right)\, dx \Big/ \int_S \exp\left(-\lambda_r F(x)\right)\, dx = \sum_{\alpha=1}^{k} \mu_\alpha z_i^\alpha, \quad i = 1, 2, \ldots, n.$$

Proof: S may be divided into compact sets S_α containing $\overline{0}_\alpha$ and disjoint except for boundaries of n-dimensional volume zero. Thus

$$S = \bigcup_\alpha S_\alpha \quad \text{and} \quad \int_S = \sum_\alpha \int_{S_\alpha}$$

for integrals with *n-dimensional volume measure.*

Let $G(x, \lambda) = \exp(-\lambda F(x))$. Then

$$(1) \qquad \int_S x_i\, G(x,\lambda)\, dx \Big/ \int_S G(x,\lambda)\, dx = \sum_\alpha \int_{S_\alpha} x_i\, G(x,\lambda)\, dx \Big/ \int_S G(x,\lambda)\, dx$$

Consider S_α in which z^α is the <u>only</u> global minimum.

Now

$$(2) \qquad \int_{S_\alpha} x_i\, G(x,\lambda)\, dx = z_i^\alpha \int_{S_\alpha} G(x,\lambda)\, dx + \int_{S_\alpha} (x_i - z_i^\alpha)\, G(x,\lambda)\, dx$$

and

$$(2.1) \qquad \left[\int_S G(x,\lambda)\, dx\right]^{-1} \int_{S_\alpha} x_i\, G(x,\lambda)\, dx = z_i^\alpha \mu(\alpha,\lambda) + \left[\int_S G(x,\lambda)\, dx\right]^{-1} \int_{S_\alpha} (x_i - z_i^\alpha)\, G(x,\lambda)\, dx$$

where $\mu(\alpha,\lambda) = \int_{S_\alpha} G(x,\lambda)\, dx \Big/ \sum_\beta \int_{S_\beta} G(x,\lambda)\, dx$.

Evidently $0 < \mu(\alpha,\lambda)$ and $\sum_\alpha \mu(\alpha,\lambda) = 1$ for all λ.

We break the integration over S_α in the last term into two parts which go to zero as the first part volume does and as $\lambda \to \infty$ in the second part.

Let

$N_\epsilon^\alpha = \{x : |x - z^\alpha| < \epsilon\}$ for $\epsilon > 0$, $\epsilon < \bar{\rho}/3$ where $\bar{\rho}$ is the minimum distance between the z^α of S.

Then

$$(3) \qquad \left| \int_{N_\epsilon^\alpha} (x_i - z_i^\alpha)\, G(x,\lambda)\, dx \right| \le \int_{N_\epsilon^\alpha} |x_i - z_i^\alpha|\, G(x,\lambda)\, dx \le \epsilon \int_{N_\epsilon^\alpha} G(x,\lambda)\, dx$$

so that

$$(3.1) \qquad \left| \left[\int_S G(x,\lambda)\, dx \right]^{-1} \int_{N_\epsilon^\alpha} (x_i - z_i^\alpha)\, G(x,\lambda)\, dx \right| \le \epsilon$$

On $S_\alpha - N_\epsilon^\alpha$, which is compact again, the minimum of the continuous function $F(x) - F(z^\alpha) \ge 0$ is, say, $\delta > 0$, since z^α is the only global minimum of $F(x)$ in S_α.

Let $M_\alpha = \max |x - z^\alpha|$ for $x \in S_\alpha$. Then, multiplying numerator and denominator by $\exp(\lambda F(z^\alpha))$,

$$(4) \qquad \left| \left[\int_S G(x,\lambda)\, dx \right]^{-1} \int_{S_\alpha - N_\epsilon^\alpha} (x_i - z_i^\alpha)\, G(x,\lambda)\, dx \right| \le M_\alpha \left[\int_S H(x,\lambda)\, dx \right]^{-1} \int_{S_\alpha - N_\epsilon^\alpha} H(x,\lambda)\, dx$$

where $H(x,\lambda) = \exp(\lambda F(z^\alpha))\, G(x,\lambda) = \exp[-\lambda \overline{F(x) - F(z^\alpha)}]$

$$(4.1) \qquad \le \exp(-\lambda\delta) \quad \text{on} \quad S_\alpha - N_\epsilon^\alpha.$$

so that

$$\int_{S_\alpha - N_\epsilon^\alpha} H(x,\lambda)\,dx \le \exp(-\lambda\delta)\; V(S_\alpha) \tag{4.2}$$

where $V(S^\alpha)$ is the volume of S_α.

Since $F(x) - F(z^\alpha)$ is continuous, for $x \in N_\eta^\alpha = \{x : |x - z^\alpha| < \eta\}$,

$F(x) - F(z^\alpha) < \delta/2$. Hence

$$\int_S H(x,\lambda)\,dx \ge \int_{N_\eta^\alpha} H(x,\lambda)\,dx \ge V(N_\eta^\alpha)\exp(-\lambda\delta/2) \tag{4.3}$$

and (4) is bounded above by

$$M_\alpha\, V(S_\alpha)\exp(-\lambda\delta/2) \Big/ V(N_\eta^\alpha) \tag{4.4}$$

Thus

$$\left| \left[\int_S G(x,\lambda)\,dx\right]^{-1} \int_{S_\alpha} x_i\, G(x,\lambda)\,dx - z_i^\alpha\,\mu(\alpha,\lambda) \right| \le \epsilon + \exp(-\lambda\delta/2)\, M_\alpha\, V(S_\alpha) \Big/ V(N_\eta^\alpha) \tag{5}$$

which goes to zero as $\epsilon \to 0$ and $\lambda \to \infty$.

Since $0 < \mu(\alpha,\lambda)$ and $\sum_\alpha \mu(\alpha,\lambda) = 1$ for all λ, every sequence $\{\lambda_n\} \to \infty$ contains a subsequence for which $\mu(\alpha,\lambda_r)$ converges to some $\mu(\alpha)$, $0 \le \mu(\alpha)$, $\sum_\alpha \mu(\alpha) = 1$.

Note that $\mu(\alpha,\lambda)$ and $\mu(\alpha)$ are independent of the coordinate x_i.

Therefore, going back to (1), every non-negative sequence $\{\lambda_r\} \to \infty$ contains (b.a.o.n.) a subsequence $\{\lambda_r\}$ such that

$$(6) \qquad \lim_{\lambda_r \to \infty} \int_S x_i \, G(x, \lambda) \, dx \Big/ \int_S G(x, \lambda) \, dx = \sum_{\alpha=1}^{k} \mu(\alpha) \, z_i^{\alpha}$$

Q.E.D.

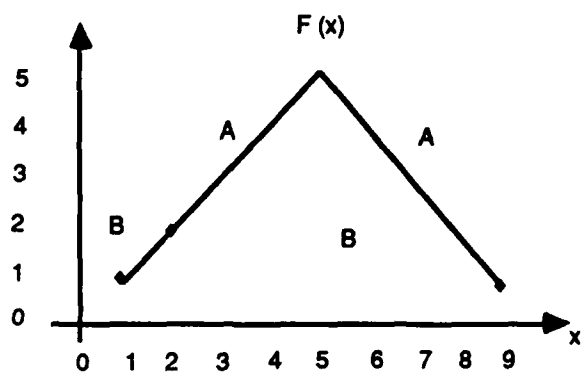
In Pincus' formula, $k = 1$, $\mu(\alpha, \lambda) \equiv 1$, so that no subsequence is necessary.

The restriction to a subsequence of $\lambda \to \infty$ can be removed sometimes by additional special conditions on F(x) and the properties of S in the neighborhood of the z^{α}, conditions which may be fulfilled in some optimization applications but not ones with F(x) non-convex at a z^{α}.

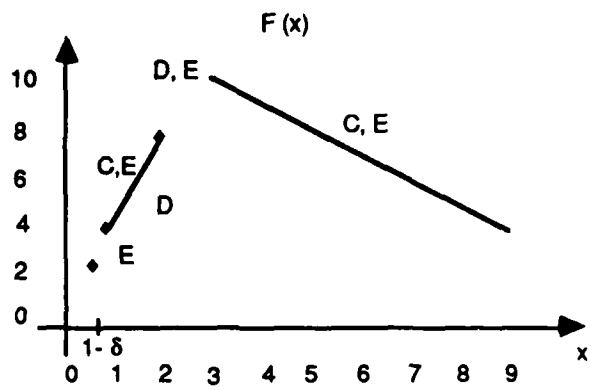
Some Examples

First, however, to give some feeling for the convergence situation, we consider several one-dimensional explicitly integrable examples as in Figure 1:

FIGURE 1



Convergence: A, B ~ 5



Convergence: C, D ~ 8.77, E ~ 1– δ

A: $F(x) = \begin{cases} x, & 1 \le x \le 5 \\ 10-x, & 5 \le x \le 9 \end{cases}$

Here $S_1 = \{x : 1 \le x \le 5\}$ and $S_2 = \{x : 5 \le x \le 10\}$

(A.1) $\int_{S_1} x\, G(x, \lambda)\, dx = \lambda^{-1} \left[e^{-\lambda} + 0\,(e^{-5\lambda})\right]$; $\int_{S_1} G(x, \lambda)\, dx = \lambda^{-1} \left[e^{-\lambda} + 0\,(e^{-5\lambda})\right]$

verifying $z^1 = 1$.

(A.2) $\int_{S_2} x\, G(x, \lambda)\, dx = \lambda^{-1} \left[9e^{-\lambda} + 0\,(e^{-5\lambda})\right]$; $\int_{S_2} G(x, \lambda)\, dx = \lambda^{-1} \left[e^{-\lambda} + 0\,(e^{-5\lambda})\right]$

verifying $z^2 = 9$.

Clearly, for large λ, $\mu(1, \lambda) = 1/2 + 0(e^{-4\lambda})$ and $\mu(2, \lambda) = 1/2 + 0(e^{-4\lambda})$

Thus, as $\lambda \to \infty$, convergence of (1) is to

(A.3) $\frac{1}{2}(1) + \frac{1}{2}(9) = 5$, the global **maximum** of F(x)!

Next consider

(B.1) $F(x) = \begin{cases} x, & 1 \le x \le 2 \\ (16-x)/7, & 2 \le x \le 9 \end{cases}$

Then $\mu(1, \lambda) = 9/16 + 0(e^{-2\lambda})$; $\mu(2, \lambda) = 7/16 + 0(e^{-2\lambda})$ and convergence is to

(B.2) $(9/16)\cdot 1 + (7/16)\cdot 9 = 5$

The value here is 11/7, over 50% higher than the minimum of 1. Note that the difference in shape and behavior of F(x) in S_1 and S_2 has made $\mu(1) \neq \mu(2)$.

(C) $F(x) = \begin{cases} 4x, & 1 \le x \le 2 \\ 13-x, & 3 \le x \le 9 \end{cases}$

Here $S_1 = \{x : 1 \le x \le 2\}$, $S_2 = \{x : 3 \le x \le 9\}$. S is not convex and not connected.

In obvious abbreviation,

(C.1) $\int_{S_1} x\, G dx = (4\lambda)^{-1} \left[e^{-4\lambda} + 0\,(e^{-8\lambda})\right]$; $\int_{S_1} G dx = (4\lambda)^{-1} \left[e^{-4\lambda} + 0\,(e^{-8\lambda})\right]$

verifying $z^1 = 1$.

(C.2) $\int_{S_2} x\, G dx = 9\lambda^{-1} e^{-4\lambda} + 0\,(\lambda^{-1} e^{-10\lambda})$; $\int_{S_2} G dx = (\lambda)^{-1} \left[e^{-4\lambda} + 0\,(e^{-10\lambda})\right]$

verifying $z^2 = 9$.

For large λ, $\mu(1, \lambda) \cong 1/37 + 0(e^{-4\lambda})$, $\mu(2, \lambda) = 36/37 + 0(e^{-6\lambda})$ and convergence is to

$$1 \cdot (1/37) + 9 \cdot (36/37) \cong 8.77$$

The sharper rate of rise of F(x) and much smaller volume in S_1 re S_2 has nearly eliminated the S_1 contribution.

D: Extending F(x) through $2 \le x \le 3$ by $F(x) = 4 + 2x$ so that F is continuous and S convex and connected, the contribution as $\lambda \to \infty$ is $0\,(\lambda^{-1}\ e^{-10\lambda})$ in $\int_{S^1} x\, G\, dx$ and is $0\,(\lambda^{-1}\ e^{-8\lambda})$ in $\int_{S^1} G\, dx$ so that convergence is to the same values as in example C.

E: Extend F(x) to $F(x) = \begin{cases} 4x, & 1-\delta \le x \le 2 \\ 4 + 2x, & 2 \le x \le 3 \\ 13-x, & 3 \le x \le 9 \end{cases}$ to have a unique global minimum at $x = 1-\delta$, with

$S_1 = \{x: 1-\delta \le x \le 3\}$ and $S_2 = \{x: 3 \le x \le 9\}$ and value $4 - 4\delta$ for $0 < \delta$ very small.

Then

$$(E.1) \quad \int_{S_1} x\,G\,dx \;/\; \int_{S} G\,dx = \left[(1-\delta) + 0\,(e^{-4(1-\delta)\lambda}\right] / \left[1 + 4e^{-\delta\lambda} + 0\,(e^{-\delta\lambda})\right]$$

and

$$(E.2) \quad \int_{S_2} x\,G\,dx \;/\; \int_{S} G\,dx = 36\,e^{-4\delta\lambda}\,[1 - 4e^{-4\delta\lambda}]$$

Thus for small δ, it is only for large λ that the unique and very little smaller global minimum is picked up e.g. that (E.2) $\rightarrow 0$.

Special and Discrete Pincus Theorem Extensions

Under special conditions on F(x) and S, asymptotic expansions for the integrals of $x_i\,G(x, \lambda)$ and $G(x, \lambda)$ for large λ may be obtained by Laplace's method, see pages 36-37 of Erdelyi's monograph "Asymptotic Expansions" [7] and by Hsu's extensions to multiple integrals [8], [9]. We use only the $G(x,\lambda) = \exp[-\lambda F(x)]$ cases, since it is only these in which we might need to consider a subsequence of λ tending to infinity. Thus, for us, Hsu's results, pages 629-30 of [9] for unique global minimum on the boundary of S_α or page 626 of [8] for interior global minimum of F(x) in S_α, would give us

$$\lim_{\lambda \rightarrow \infty} \int_{S_\alpha} \exp[-\lambda F(x)]\,dx = \exp[-F(z^\alpha)] \left\{ (2\pi)^n / H_n[-F(z^\alpha)] \right\}^{\frac{1}{2}},$$

where $H_n\,[F(z^\alpha)]$ is the n-dimensional Hessian of F at z^α. The set S_α is supposed to contain a closed finitely connected domain with continuously turning targent hyperplane which includes z^α. The function of F is to be of class C^2 (continuous with continuous first and second partial derivatives). The Hessian required to be positive implies that F(x) is strictly convex at z^α, a most

restrictive condition. Thus, Pincus' result is much stronger than the asymptotic Laplace-Hsu results.

Discrete Case

Consider a bounded discrete set of points $D \subseteq E^n$ and a function $F(x)$ defined on D and with global minimum there at z^α, $\alpha = 1, \ldots, k$.

Theorem 2: If

(1) $F(x)$ has a finite number k of global minima in D at z^α, $\alpha = 1, \ldots, k$.

(2) $|D| \leq V_D$ and for some $\lambda_0 > 0$,

$$\sum_{x \in D} \exp\ [-\lambda_0 F(x)] < \infty$$

(3) $$\inf_{D \setminus \bigcup_\alpha (z^\alpha)} [F(x) - F(z^\alpha)] = \delta > 0$$

Then

$$\lim_{\lambda \to \infty} \sum_{x \in D} x_i \exp[-\lambda F(x)] \Big/ \sum_{x \in D} \exp[-\lambda F(x)] = k^{-1} \sum_{\alpha=1}^{k} z_i^\alpha, \quad i = 1, \ldots, n$$

Proof: Let $G(x, \lambda) = \exp[-\lambda F(z^\alpha)]$, $H(x, \lambda) = \exp[-\lambda \overline{F(x) - F(z^\alpha)}]$,

$Z = \bigcup_\alpha \{z^\alpha\}$ and $D \setminus Z = D'$.

Then

$$(7) \quad \sum_{x \in D} x_i\, G(x, \lambda) \Big/ \sum_{x \in D} G(x, \lambda) = \sum_{x \in D} x_i\, H(x, \lambda) \Big/ \sum_{x \in D} H(x, \lambda)$$

$$(7.1)\quad = \left[\sum_{\alpha=1}^{k} z_i^{\alpha} + \sum_{x\in D'} x_i\, H(x,\lambda)\right] \Big/ \left[k + \sum_{x\in D'} H(x,\lambda)\right]$$

Let $\lambda = \lambda_0 + \lambda_1$, $\lambda_1 \geq 0$. Then

$$(7.2)\quad H(x,\lambda) = H(x,\lambda_0)\ \exp\left[-\lambda_1 \overline{F(x) - F(z^{\alpha})}\right] \leq \exp(-\lambda_1\, \delta)\, H(x,\lambda_0)$$

So

$$(8)\quad \sum_{x\in D'} H(x,\lambda) \leq \exp(-\lambda_1\, \delta) \sum_{x\in D'} H(x,\lambda_0)$$

$$(8.1)\quad \sum_{x\in D'} x_i\, H(x,\lambda) \leq V_D \sum_{x\in D'} H(x,\lambda_0) \leq V_D \exp(-\lambda_1\, \delta) \sum_{x\in D'} H(x,\lambda_0)$$

$$(8.2)\quad \lim_{\lambda\to\infty} \sum_{x\in D'} H(x,\lambda) = \lim_{\lambda\to\infty} \sum_{x\in D'} x_i\, H(x,\lambda) = 0$$

and from (7.1),

$$(9)\quad \sum_{x\in D} x_i\, G(x,\lambda) \Big/ \sum_{x\in D} G(x,\lambda) = k^{-1} \sum_{\alpha=1}^{k} z_i^{\alpha},\quad i = 1,\ldots,n.$$

Q. E. D.

From Optimization to Simulated Annealing

Pincus, understanding that his 1968 results [1] could not be made operational without a means of approximately evaluating integrals over irregular sets and in multi-dimensional spaces, provided in 1970 [2] a method of computation. Since F(x) was continuous on a compact set S, the Riemann integrals could be approximated by summing the values of the integrands at a finite grid of points in S each multiplied by the volume of its grid cell. His key was the World War II work of von Neumann, Ulam, Everett and Metropolis [3] as partly summarized but not referenced in

Chapter 9 of the excellent 1964 Methuen monograph of Hammersley and Handscomb, Monte Carlo Methods [10].

An incisively clear rendition of the backgrounds and basic idea of associating an ergodic irreducible (finite state) Markov chain satisfying the strong law of large numbers so that a sample sequence of Markov transitions could give with probability one an expected value which would be a desired integral can be found in "The Monte Carlo Method" by N. Metropolis and N. Ulam in the 1949 Journal of the American Statistical Association, pp. 335-341, [13]. This paper which references other von Neumann, Everett, Ulam work makes clear (page 338) that the method had been and was being applied to many problems involving (partially) stochastic flows governed by integro-differential equations, not just to Boltzmann equilibria in "statistical mechanics." Ignorance or unavailability of these references likely led to the current misapprehension of history as in the preface to the 1987 book [11] of van Laarhoven and Aarts Simulated Annealing: Theory and Applications, D. Reidel Publishing Co., Dordrecht, Holland.

Pincus' work dealt with optimization problems with a unique optimum. His expressions for this optimum as a limit of expected value expressions led him to make the Monte Carlo Markov chain connection. Our results for non-unique global optima, the usual case, provides insight into convergence properties of associated Markov sample runs designated "simulated annealing," free of metaphysical non-mathematical ideas of melting, cooling and freezing.

Markov Chain Computation and Simulated Annealing

Pincus in [2], borrowing from [3], associated with each point x^k of S in the approximating Riemann integral sums

$$(10.1)\quad \sum_{k=1}^{N} x_i^k \exp\left[-\lambda F(x^k)\right] \quad \text{and} \quad \sum_{k=1}^{N} \exp\left[-\lambda F(x^k)\right]$$

The N-vector of coordinates (for all $\lambda > 0$)

$$(10.2)\quad \Pi_j = \exp\left[-\lambda F_j\right] \Big/ \sum_k \exp\left[-\lambda F_k\right] \quad , \quad j = 1, \ldots, N$$

where $F_k \equiv F(x^k)$ and λ is omitted where clear from context. The $\Pi_j > 0$ are to be the (unique) "invariant" or "stationary" distribution of an (ergodic) irreducible, aperiodic Markov process with N "states" (or "configurations" in simulated annealing "SA" nomenclature) and matrix (P_{ij}), $i, j = 1, \ldots, N$. Starting with an arbitrary symmetric, irreducible, aperiodic Markov matrix P^*_{ij}, P_{ij} is determined by

$$(11)\quad P_{ij} = \begin{cases} P^*_{ij}\, \Pi_j / \Pi_i \ , \text{ if } \Pi_j / \Pi_i < 1, \ i \neq j \\ P^*_{ij} \qquad\qquad , \text{ if } \Pi_j / \Pi_i \geq 1, \ i \neq j \\ P^*_{i\,i} + \sum_{j \in J^-} P^*_{i\,j}\left(1 - \Pi_j / \Pi_i\right), \quad i = j, \ J^- = \left\{j : \Pi_j / \Pi_i < 1\right\} \end{cases}$$

The P^*_{ij} are designated G_{ij} ("generation probabilities") in SA literature which has $P_{ij} = G_{ij} A_{ij}$ where A_{ij} is an "acceptance" probability matrix, see [11], page 13. The SA literature also uses $\lambda^{-1} = c$, or T, so that as $\lambda \to \infty$, c, or T (termperature) $\to 0$.

The Markov (sample) chains are defined as follows:

If currently in "state" i, use the P^*_{ij} (G_{ij}) probabilities to pick state j. Calculate Π_j / Π_i. If $\Pi_j / \Pi_i \geq 1$, accept j as the new state. If $\Pi_j / \Pi_i < 1$, then with probability Π_j / Π_i accept j and with $1 - \Pi_j / \Pi_i$ go back to the original state i. This procedure corresponds to sample chains using P_{ij}. Operationally, $\Pi_j / \Pi_i \geq 1$ means $F_j \leq F_i$; $\Pi_j / \Pi_i < 1$ means $F_j > F_i$. Acceptance of j is (Metropolis criteria) by comparing $\exp[-\lambda F_j]$ with a random number R drawn from the uniform distribution on (0,1]. Accept if $\exp[-\lambda F_j] > R$, otherwise stay at i.

How long a Markov chain to use for each λ value is not well specified in SA, see [11] pages 59-61, but the idea is (page 10) "until equilibrium is approached sufficiently closely", whatever "equilibrium", undefined mathematically, may mean. Supposing it means getting close to optimum, this is usually impossible to estimate or else because of horrendous to do computation time. Pincus in [2] gives a rough Chebychev inequality bound re convegence to the expected value for the case of a **unique** global minimum. A similar bound is mentioned [11] page 56 without reference to Pincus and for, presumably, the general multiple optima situation which Hajek [5] and others explicitly mention. All [11, 12] seem to have the idea that one has "convergence" to a single state i.e. that a unique minimum state for each λ^{-1} ("temperature") employed is achieved in computation which is stopped either when the minimum state repeats itself often or by choosing the state reached when a fixed number of transitions is reached. Proof that such repetition can "unreasonably" often occur (even for simple heads-tails coin tossing) may be found in Feller [13] Chapter III, especially pages 77–86, re the Arc Sine law and Probabilities of Long Leads.

Example Computations

Nearly all SA computation (and much theory) has employed P^*_{ij} matrices with equal P^*_{ij}'s in a band around the main diagonal e.g., uniform non-zero probabilities between "neighboring" states, zero between non-neighboring states but with positive probability of reaching any one state from any other in possibly multiple steps (the 'irreducibility" property).

Abstracting 3 states, 1, 2, 3, corresponding to x = 1, 5, 9 from example A with F_1= 1, F_2= 5, .F_3= 1, using the uniform (P^*_{ij}) matrix of elements P^*_{ij} = 1/3, choosing λ = 0.5 (temperature = 2) and making a sample run of 96 transitions starting with state 1, we reached state one 47 times, state two 7 times, state three 42 times. No sojourn in a state lasted more than 4 times; every state was a global minimum starting with transition 66 whereupon the transitions were flip-flops between the global minimum states 1 and 3 rather than an "equilibrium" at either. The

relative frequencies of 0.4948, 0.0722, 0.433 for respectively states one, two, three correspond to the theoretical values of 1/2, 0, 1/2 from our "discrete" Pincus Theorem.

We next consider a class of travelling salesmen problems with 2n cities having obvious global optimal tours of minimum distance traveled. The city locations are at points (1, 1), . . . , (1, n), (n, 0), . . . , (1, 0) in the plane with "rectangular" (e.g. "l_1") distance between pairs of cities. We label the tours (or configurations) as cyclic permutations (1, . . . , 2n) where (1, 1), . . . (1, n), respectivety correspond to cities 1, . . . , n and (n, 0), . . . , (1, 0) correspond to n + 1, . . . , 2n. Clearly the global optimal tours are those starting at any city and going clockwise (or counter clockwise) around a "rubberband" encircling the cities. The transitions are an exchange of a pair of cities in a permutation. Thus

(a) There are 4n global optimal tours

(b) (2n)! tours, with many local minima

(c) The a priori probability of finding a global optimum is quite small, 0 (4n / (2n) !)

(d) The average tour length is 0(n^2)

(e) The optimal tour length is 2n < < 0 (n^2).

Starting with 2n = 500, the Kirkpatrick et al "Metropolis" algorithm [4], the tour (1, . . . , n, 2n, 2n-1, . . . , n + 1) , we made a number of runs on a SEQUENT computer (roughly equivalent to a VAX 780). We present results for some "high" temperature runs using 1000 transitions and "low" temperature runs using 3000 transitions - - - 1000 transitions took 45 CPU minutes on the computer.

Notice that at high temperatures the "equilibrium" tends toward the average case; at low temperatures it gets stuck in local minima.

Simulated Annealing Runs

(500 Cities, 1,000 Transitions per run, starting tour cost 998, optimal cost 500.)

	Temperature (λ^{-1})	Cost Last Tour	Cost Best Tour Found
	10.0000	4494	998
(Case 1)	1.0000	3284	998
	0.1000	3260	998

(3,000 transitions per run, starting tour cost 998)

	1.0000	1146	998
	0.1000	1014	998
(Case 1)	0.5000	1016	998
	0.1000	998	998
	1.0000	1146	998
(Case 2)	0.1000	1014	998
	0.0100	1014	998
	1.0000	1150	998
(Case 3)	0.1000	1040	998
	0.0100	1040	998
	1.0000	1128	998
(Case 4)	0.5000	1024	998
	0.1000	1010	998
	0.0100	1010	998
	1.0000	1206	998
(Case 5)	0.1000	1026	998
	0.0100	1026	998

Individual Sample Runs Versus Ensemble Properties

Simulated annealing analogies to physical or thermal equilibrium statistical mechanics confuse between average properties of ensembles, properties of averaged behavior of many individual particles, with that of individual particle behavior. The vast difference mathematically between ensemble properties and behavior in individual sample runs is brought out in Feller [13] Chapter III, "Coin Tossing and Random Walks." Without such behavior there would likely be no gamblers nor attempts at gambling systems.

In our first SA example we chose only three states and 96 transitions in order to ensure that the ergodicity property of the sample Markov chain would take over. As mentioned, the latter two-thirds of the transitions involved flip flops between the two global minima states (with no visitation of the poorer state). Is this equilibrium, the state at which to terminate the sample run, as undefined in von Laarhoven and Aarts [11] especially page 10?

In the second example, starting with a state close to optimum in the sense that only 2 particular interchanges would suffice to obtain an optimum, not only did we fail to improve from this state but instead got substantially worse results for the higher (melting) temperatures. Of course, we only (!) did 1,000 and 3,000 transitions for a problem in which 500! transitions are *a priori* possible. But 45 minutes (CPU time) per 1,000 transitions were required on a mainframe computer. And, "equilibrium" seemed to be attained on these runs.

Many combinatorial optimization problems can be stated in continuum form and solved,at worst approximately, by continuum methods. For example, integer programming problems with network constraints and integer data can be solved exactly with extreme point algorithms which are also two orders of magnitude more efficient in time and size accomodated than general purpose LP algorithms. Christophides algorithm for the traveling salesman problem based on network structure can guarantee no worse than 150% over optimal cost. The simulated annealing procedures can guarantee nothing.

Another point exhibited by our Pincus examples is that the convergence of the integral formulations may be to a different convex combination of the global minima from that of the approximating Riemann sums. The latter, via our Discrete Case Theorem, is always to the simple average of the global minima. Thus the interchange of limit operations, $\lambda \rightarrow \infty$ and sum $\rightarrow$ integral can make large differences in the fluctuation of states visited (and how often)

within a sample run. This point applies whether or not the optimum is unique since approximate non-global optima may have equal values. Sample runs often converge, see Hammersley and Handscomb [10], to local non-global optima (with flip flops) since transitions to better states may be possible only through (very unlikely) transitions to states poorer than the local optima.

Thus simulated annealing computations cannot be trusted to deliver global optima.

References

[1] M. Pincus, "A Closed Form Solution of Certain Programming Problems", Operations Research, (1968) 16, 690-694.

[2] M. Pincus, "A Monte Carlo Method for the Approximate Solution of Certain Types of Constrained Optimization Problems", Operations Research, (1970) 18, 1225-1228.

[3] N. Metropolis and S. Ulam, "The Monte Carlo Method", Journal of the American Statistical Association , (1949) 44, 335-341.

[4] S. Kirkpatrick, C.D. Gelatt, Jr., M.P. Vecchi, "Optimization by Simulated Annealing", Science , (1983) v. 220, no. 4598, 671-680.

[5] B. Hajek, "Cooling Schedules for Optimal Annealing", Mathematics of Operations Research , (1988) v. 13, no. 2, 311-329.

[6] T. Chiang and Y. Chow, "On Eigenvalues and Annealing Rates", Mathematics of Operations Research, (1988) v. 13, no. 3, 508-511.

[7] A. Erdelyi, Asymptotic Expansions, Dover Publications, Inc. (1956).

[8] L. C. Hsu, "A Theorem on the Asymptotic Behavior of a Multiple Integral", American Journal of Mathematics , (1948), 623–632.

[9] L. C. Hsu, "On the Asymptotic Evaluation of a Class of Multiple Integrals Involving a Parameter", American Journal of Mathematics , (1951), 625–634.

[10] J.M. Hammersley and D.C. Handscomb, Monte Carlo Methods, Methuen's Monographs on Applied Probability and Statistics, London, (1964).

[11] P.J.M. van Laarhoven and E.H.L. Aarts, Simulated Annealing: Theory and Applications, D. Reidel Publishing Co., Dordrecht, Holland (1987).

[12] R.A. Rutenbar, "Simulated Annealing Algorithms: An Overview", IEEE Circuits and Devices Magazine, January (1989), 19-26.

[13] W. Feller, An Introduction to Probability Theory and its Applications, Volume 1, 2nd Edition, J. Wiley and Sons (1957).

ADA207537

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER CCS 615	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Extended Pincus Theorems and Convergence of Simulated Annealing		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER CCS 615
7. AUTHOR(s) A. Charnes M. Wolfe		8. CONTRACT OR GRANT NUMBER(s) SES 8520806 SES 8722504 DAKF 15-87-C-0110
9. PERFORMING ORGANIZATION NAME AND ADDRESS Center for Cybernetic Studies, UT Austin Austin, Texas 78712		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research, Washington, D.C. National Science Foundation		12. REPORT DATE December 1988 13. NUMBER OF PAGES 21
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Extended Pincus Theorems, Global Optimization, Simulated Annealing, and Ergodic Markov Chains

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Pincus' 1968 formula for the (unique) global minimum of a continuous function on a compact set in E^n is extended to finite multiple optima and to discrete and special variants. The impact of these on associated erodic irreducible aperiodic Markov chain computation he initiated, (1970) currently called "simulated annealing", is exemplified and assessed leading to grave concern about what current simulated annealing processes may converge to instead of optima.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601 |

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

END

FILMED

6-89

DTIC